Combination of deep learning and syntactical approaches for the interpretation of interactions between text-lines and tabular structures in handwritten documents

Camille Guerry, Bertrand B. Coüasnon, Aurélie Lemaitre

Combination of deep learning and syntactical approaches for the interpretation of interactions between text-lines and tabular structures in handwritten documents

Camille Guerry, Bertrand Coüasnon, Aurelie Lemaitre
Univ Rennes, CNRS, IRISA
Rennes, France
{*camille.guerry, bertrand.couasnon, aurelie.lemaitre*}*@irisa.fr*

***Abstract*—In this article, we present our work on baseline detection in images of historical documents. This work focuses on handwritten documents containing tabular structures. One of the difficulties of this kind of documents is the strong interaction between text and tabular structures. This interaction leads to ambiguous cases for which recognition systems often over- or sub-segment baselines.**

The interest of our method is to combine contextual and structural knowledge in order to interpret properly this interaction. Our combination is able to merge heterogeneous information obtained with a deep-learning approach (for contextual elements) and a syntactical approach (for structural elements). Our grammatical description consists on a logical description of the intersections between text-lines and vertical rulings of detected tables. Intersections are described thanks to physical indicators extracted from images: vertical rulings, hypothetical text-lines, begin- and end-indicators of text-lines.

We show on cBAD competition [4] (competition on baseline detection) that the combination of heterogeneous knowledge (structural and contextual information) improves baseline detection in handwritten documents. We obtain better scores than the best method published until now on this competition.

***Keywords*-tabular structure recognition; old handwritten document processing; syntactical approach; deep learning**

I. Introduction

Baseline detection is an important step for the recognition of historical digitized documents. Indeed, many handwritten text recognition methods take cropped images of text lines as an input. To produce those cropped images, it is previously necessary to extract baselines in document images. In old handwritten documents, the presence of tabular structures leads to specific difficulties for baseline segmentation. This is due to the strong interactions between text and tabular structures that can be ambiguous for the recognition system. The ambiguity of the interaction between text and tabular structure is shown on Figure 1:

- case (a): two different lines that belong to two different columns are very close to each other,
- case (b)/(c): a single text-line crosses the tabular structure and the physical ruling of the table has no effect on it or the text-line overlaps on the nearby column of the tabular structure.

In case (a), the recognition system should produce two different baselines but in case (b)/(c) it must produce a single baseline. It is difficult to differentiate those cases without contextual information. Another difficulty of text-line segmentation in document images is due to the way documents are digitized: text belonging to adjacent pages is sometimes visible. This text must be considered as noise.

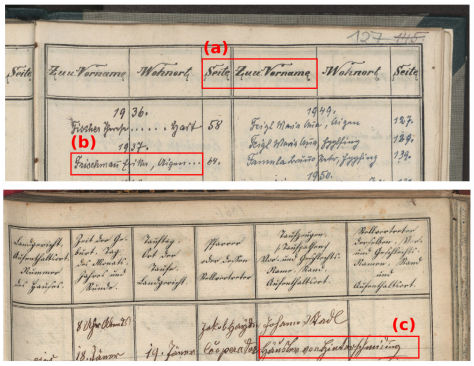

Figure 1. Examples of ambiguity generated by the interaction between text and tabular structures: (a) two different lines are very close from each other, (b) a single line crosses the tabular structure, (c) a text line exceed on the nearby column of the tabular structure.

The main contribution of our work is to propose a way to combine structural knowledge and contextual information in order to improve baseline segmentation in the context of tabular structures. One of the interests of our combination is its ability to merge heterogeneous elements obtained with a deep learning and a syntactical approach. We propose to combine:

- contextual elements obtained from different CNN outputs: hypothetical baselines, begin- and end-indicators of baselines, and page boundary (deep-learning part),
- structural indicators (rulings of table) built thanks to a grammatical description (syntactical part),
- structural knowledge on table (syntactical part).

In section II of this paper, we present works related to text-line localisation and tabular structure recognition. Section III is dedicated to the combination of contextual

information and structural knowledge that we propose for baseline segmentation in the context of tabular structures. In section IV, we test our combination on the complex dataset of cBAD competition (track B) and on subset of this dataset containing exclusively documents with tabular structures.

II. Related Work

Recently, systems based on CNNs outperform most of traditional methods on document analysis tasks, and in particular on baseline detection. Indeed, CNN-based methods obtained the most competitive results in the ICDAR2017 Competition on Baseline Detection in Archival Document (cBAD). The models were evaluated on several challenging datasets proposed by Diem *et al.* [4] and composed of documents from 9 different archives.

The approach proposed by Fink *et. al.* [3] uses two CNNs that follow a U-net architecture: one for text region detection and classification of documents depending on basic layout properties (pre-processing step) and the other for baseline extraction. Then, a post-processing step takes advantages of the basic layout properties obtained with the pre-processing step.

As the dataset and the metric of cBAD are still available, other systems were tested on this dataset after the competition. One method outperforms the results obtained during the competition [9][1]. This method (dhSegment [9]) is based on a CNN which follows a U-net [10] architecture with residual blocks and uses pretrained weights learned on ImageNet [2]. He *et al.* [6] introduced the notion of residual blocks. The output of this neural network is a probability map. Oliveira *et. al.* [9] proposed a simple post-processing step based on filters in order to extract baselines from probability maps. The main advantages of dhSegment is its genericity: it has been used to solve different challenges relative to old document processing, including page extraction, baseline detection, document layout analysis, and ornament detection.

One of the two datasets proposed during the cBAD competition [4] is composed of documents with heterogeneous and complex structures (cBAD track B). This dataset notably contains documents with tabular structures but not only. cBAD metric takes into account the segmentation of baselines depending on the tabular structure.

Oliveira *et. al.* do not directly consider tabular structures but their performance suggests that the network is able to learn the interaction between text and tabular structures partly. Nevertheless, baselines obtained by Oliveira *et. al.* [9] are sub-segmented when documents contain tabular structures (see Figure 2). Oliveira et al. themselves notice subsegmentation of text-lines.

The method proposed by Lemaitre *et. al.* [7] for the cBAD competition is based on text-lines extraction in blurred images followed by a grammatical description in EPF using the DMOS-PI method.

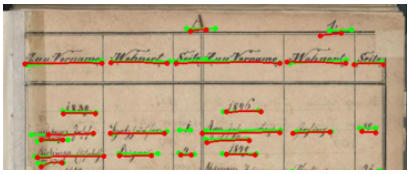

Figure 2. Results obtained with dhSegment : The ground-truth and predicted baselines are displayed in green and red respectively- the figure comes from [9]

Compared to text-line extraction in blurred images, methods based on neural networks are more able to take into account contextual information like the shape of the writing. The advantage of the method proposed by Lemaitre *et. al.* [7] is that rules on tabular structures are explicitly formulated thanks to a grammatical description. However, this method lacks contextual information to differentiate between a text-line which crosses a tabular structure and two close text-lines (see Figure 1).

As our aim is to interpret the interaction between text and tabular structures, using structural rules is especially interesting. Therefore, we chose a syntactical approach to provide structural knowledge to our system. For our structural description, we need some contextual information: hypothetical baselines, begin- and end-indicators of baseline. We decided to predict those elements using a deep-learning approach to take advantages of the high performance recently obtained by such approaches. To predict all those elements, we need a neural network architecture that is generic enough. Because of the genericity of the approach proposed by Oliveira *et. al.* [9], we choose to use the dhSegment architecture to predict all the contextual elements listed above.

III. Proposed method

A. Overview of the system

Figure 3 presents an overview of our system. We propose a method that combines contextual information obtained thanks to a deep-learning approach with structural information expressed in a syntactical way.

We use a deep-learning approach in order to predict several elements: hypothetical baselines (Figure 3 (4)), begin- and end-indicators of each baseline (Figure 3 (2) and (3)) and the boundary of each page (Figure 3 (5)). We also extract physical rulings of tabular structures (Figure 3 (1)) thanks to a line segment extractor based on a Kalman filter [8].

A grammatical description is then used to describe and detect tables and then combine all those previous elements in a structural way. Our grammatical description is written in the EPF formalism, which is the syntactical language of the DMOS [1] method that we use.

Section III-B and III-C are organized as follows. In section III-B, we present the chosen settings for the training of the

[1]Another method (the ARU-net proposed by Grüning *et. al.* [5]) obtains better results than [9] but it has not yet been published (available on arXiv).

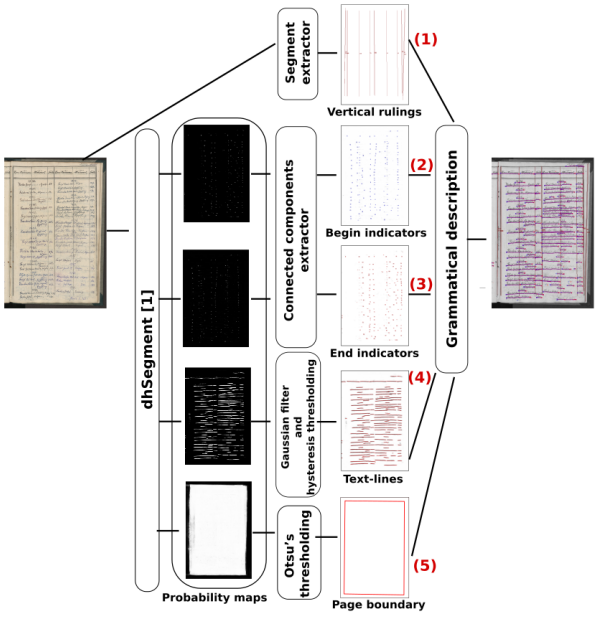

Figure 3. Overview of the system

existing deep learning approach we use: dhSegment [9]. In section III-C, we present the combination of contextual and structural knowledge that we propose for the interpretation of the interaction between text-lines and tabular structures.

B. Deep learning approach for the extraction of contextual information

We use the fully convolutional network proposed by Oliveira *et. al.* [9], without changing the architecture, in order to predict:

- baselines (Figure 3 (4));
- begin-indicators of each baseline (Figure 3 (2));
- end-indicators of each baseline (Figure 3 (3));
- page boundaries (Figure 3 (5)).

Oliveira *et. al.* already proposed baseline and page boundary prediction in [9]. We propose to train additionally dhSegment CNN for the prediction of two other classes: begin- and end-indicators of baselines. Thanks to the genericity of this system [9], it is easy to use it for the prediction of begin- and end-indicators. We perform the training of two models of dhSegment [9]. One for the prediction of baselines, begin- and end- indicators and the other for page boundaries.

The first model is trained to predict four different classes: baselines, begin-indicators, end-indicators and background. Figure 4 gives an example of an annotated image given to the model during the training phases. Baselines are annotated with a training mask of 5 pixels and extremity indicators with circles of diameter 10 pixels. Images are resized to have 10^6 pixels as proposed in [9]. The second model is trained to predict page boundaries and we reuse the settings proposed by Oliveira *et. al.* [9].

The training set used for the first model is the one proposed for the competition cBAD for the complex track

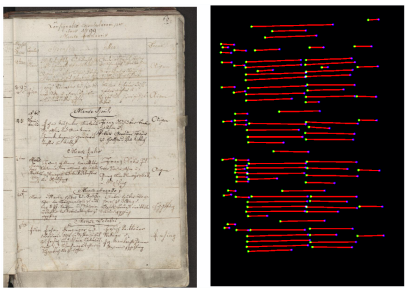

Figure 4. Image of cBAD training set and it's corresponding annotation: baselines (red), begin indicators (green), end indicators (blue)

(track B). It is composed of 271 images. For the second model, we use a training set composed of 1242 images. This training set contains all images of cBAD simple track (track A) and the images of the train set of cBAD complex track (track B).

dhSegment produces a probability map for each class. The begin- and end-indicators are extracted from the probability maps (Figure 3 (2) and (4)) thanks to a connected component extractor. For the extraction of baselines and page boundaries, we use the post-processing proposed by Oliveira et al. [9], that consists of a Gaussian filter followed by a hysteresis thresholding and Otsu's thresholding respectively.

C. Syntactical approach for the integration of structural information

*1) **Description of the interaction between text and tabular structure**:* The syntactical part of our method consists in a combination of structural knowledge and heterogeneous contextual information obtained with deep learning thanks to a grammatical description of the tabular structure.

To write our grammatical description, we need another element: vertical rulings of tabular structure. To extract vertical rulings, we use a line segment extractor based on Kalman filter [8]. As we do not have a ground-truth for vertical rulings and because obtaining this ground truth could be quite laborious, we prefer a more classical document processing method for this task.

All the extracted elements (hypothetical baselines, begin-indicators, end-indicators and vertical rulings) are used as terminal for our grammar. Our grammatical description relies on the following structural information:

- a vertical ruling often represents a logical separator for text-lines,
- an end-indicator followed by a begin-indicator also represents a separator for text-lines.

Considering only vertical rulings is not enough because of the ambiguity of the interaction between text and tabular structures (see Figure 1). Considering only extremity indicators is also not enough because of false positives that can produce other segmentation errors (see Figure 5). That is why we wrote a grammatical description that takes into account those two structural information.

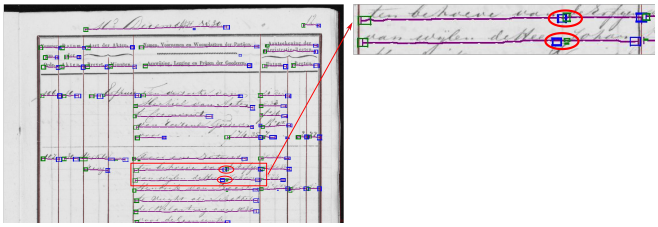

Figure 5. Elements extracted thanks to dhSegment (baselines in purple, begin-indicators in green, end-indicators in blue) - example of false positive for baselines extremity (red circle)

This grammatical description allows the detection of table structures, and for each vertical ruling of each table structure, we search all text-lines that cross it. Text-lines that cross a vertical ruling are split if and only if we find a begin-indicator followed by an end-indicator in the neighborhood of the intersection between the text-line and the ruling. Figure 6 illustrates all configurations that lead to the decision to split a baseline.

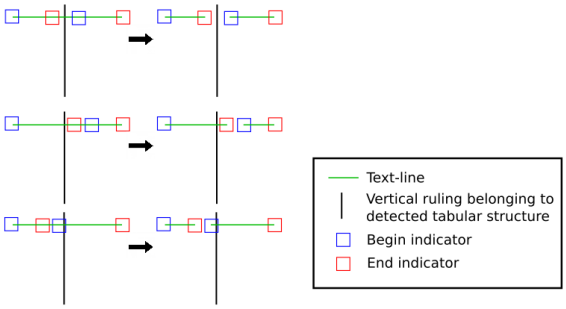

Figure 6. Illustration of combination rules. The final segmentation (on the right) depends on the position of each elements on the left.

```
intersection L1 L2 ::=
    AT(inTable)&&
    USE_LAYER(verticalRuling)FOR(aRuling R)&&

    AT(onRuling R)&&
    USE_LAYER(textLine)FOR(aLine L)&&

    AT(nearIntersection R L )&&
    USE_LAYER(endIndicator)FOR(endOfL I1)&&

    AT(rightOfIndicator L I 2)&&
    USE_LAYER(beginIndicator)FOR(beginOfL I2)&&
    splitLigne(L I1 I2 F L1 L2).
```

Figure 7. Grammatical description of an intersection between text-line and detected tabular structure

In Figure 7 we show the grammatical rule that allows us to take the decision to split a baseline. This rule is written in EPF formalism. `AT` is an operator of EPF used to search elements in a specific area of the image, modeling this way a relative position. For the `intersection` rule, we describe the following position areas: `inTable`, `onRuling`, `nearIntersection` and `rightOfIndicator`. The operator `FOR` applies the grammatical rule in argument to the perceptive layer selected by `USE_LAYER` operator. A perceptive layer stores terminal or non-terminal elements, which were previously detected. In our case `textLine`, `endIndicator` and `beginIndicator` are layers composed of terminals obtained with deep learning, while `verticalRuling` is a layer composed of the non-terminals that corresponds to vertical rulings of tables recognized thanks to another rule. `&&` is the operator of EPF that allows us to define a sequence of rules. The variables `R`, `L`, `I1` and `I2` represent respectively a vertical ruling of the table, a text-line, a begin-indicators and an end-indicator obtained thanks to the rules `aRuling`, `aLine`, `endOfL`, `beginOfL` and use as parameter for the definition of position area and for the rule that split a baseline. `L1` and `L2` are variables that contain the two baselines produced by the rule `intersection` if this rule succeeds.

*2) **Integration of the information about page boundary**:* Another information that we integrate in our combination of heterogeneous elements is the page boundaries obtained thanks to deep learning. The grammatical method we use (DMOS [1]), proposed an operator called `IN`. This operator takes as an input an area delimitation and its effect is to reduce the analysis area of the grammar. In our grammar, we use the operator `IN`, to reduce the analysis to the page area predicted by dhSegment. The operator `IN` enables to integrate information about page boundaries in a very simple and efficient way.

IV. Experimental Setup and Results

A. cBAD dataset and elaboration of a Table subset

The cBAD competition [4] on baseline detection proposes a complex dataset (track B) that notably contains tabular structures but not only. This dataset is composed of a train set (*cBAD_train_set*) containing 271 documents and a test set (*cBAD_test_set*) composed of 1010 documents.

As our method was designed for the interpretation of baseline interaction with tabular structures, we test it on a subset of *cBAD_test_set* containing exclusively documents with tabular structures, while training it on the whole *cBAD_train_set*. We call this test subset the *table_test_set*. However, identifying which structure can be considered as a tabular structure in cBAD dataset is not always obvious. This is why we consider the following rule in order to select the documents of the *table_test_set*. A document contains a tabular structure if at least one of those two properties is verified:

- the tabular structure is materialized by vertical and horizontal rulings,
- columns of the tabular structure are materialized by vertical rulings and those columns have names.

Examples of what we consider as tabular structures or not are shown in Figure 8. The subset of tables we build

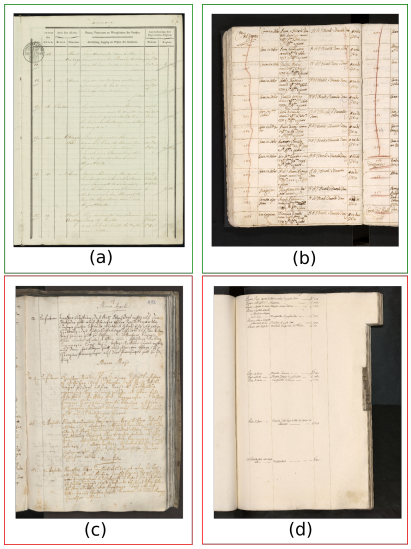

Figure 8. Examples of documents selected for the *table_test_set*: (a) and (b) - Examples of documents not selected for the *table_test_set*: (c) and (d)

(*table_test_set*) is composed of 315 documents (51 084 text-lines) out of a total of 1010 documents (88 962 text-lines). The names of the images selected from *cBAD_test_set* for our *table_test_set* are available here [2].

In order to compare our system with state-of-the-art methods, we also test it on the full complex *cBAD_test_set* (track B) composed of 1010 documents.

B. Results on Table subset

Table I
COMPARISON WITH DHSEGMENT ON *table_test_set* (315 IMAGES)

Method	P-val	R-val	F-val
dhSegment (retrained CNN)	0.870	**0.964**	0.915
Combination *(text-lines, vertical rulings, begin- and end-indicators)*	0.895	**0.964**	0.928
Combination *(text-lines, page boundary)*	0.877	0.961	0.917
Global combination	**0.901**	0.961	**0.930**

Table I summarizes results obtained on the *table_test_set*. The implementation of dhSegment [9] is available on github but the weights that the authors obtained after a training on baselines detection are not given. In order to use dhSegment CNN [9], we need to retrain it. We choose to train dhSegment CNN on the full *cBAD_train_set* given by the competition (track B, 271 images) even if the latter is not only composed of documents with tabular structures.

We test the retrained CNN of dhSegment [9] on the *table_test_set* that we built (see IV-A). On this subset dhSegment obtains a precision (P) of 0.870, a recall (R) of 0.964 and a final value (F) of 0.915. We can notice that the results on this subset are better than those on the full dataset. This is due to the higher quality of the images containing tabular structures, which are globally less degraded than other images in the cBAD dataset.

We integrate the text-lines produced by dhSegment in our global combination system. The syntactical approach of our method is divided in two different tasks, so we measure the improvement provided by:

- the combination of text-lines, vertical rulings, begin- and end-indicators for the interpretation of interactions between text and tabular structures ((1), (2), (3), (4) in Figure 3),
- the combination of text-lines and page boundary to delete false positive detected on adjacent pages ((4) and (5) in Figure 3),
- the global combination: text-lines, vertical rulings, begin- and end-indicators and page boundary.

Thanks to the interpretation of the interaction between text and tabular structures, we increase the precision score from 87.0% to 89.5% and the F-value from 91.5% to 92.8%. As expected, our method has no effect on the recall score because its aim is to cut baselines correctly depending on a contextual and structural context: we do not try to detect missing baselines. The integration of page boundary in the global system also improves the precision score. However, we lose 0.3% on the recall score. This loss is produced by bad page detection. Indeed, when the page detected by the neural network is smaller than the real page, this leads to removal of some correct baselines. Thanks to the global combination ((1)+(2)+(3)+(4)+(5)), we increase the precision score from 87.0% to **90.1%** and the F-value from 91.5% to **93.0%**.

In Figure 9 we show some examples of improvement obtained thanks to our combination. In some cases, our combination produces errors (see Figure 9). However, we can notice that the right line segmentation in those cases is difficult to determine even for a human.

C. Comparison with state of the art on cBAD whole dataset

We showed in the previous section that our combination method is able to improve baseline detection in a subset of cBAD composed of documents with tabular structures (*table_test_set*). Consequently, our method also improves baseline detection in the full *cBAD_test_set* (track B).

Table II
RESULTS ON *cBAD_test_set* (TRACK B, 1010 IMAGES)

Method	P-val	R-val	F-val
UPVLC [4]	0.833	0.606	0.702
IRISA [4]	0.692	0.772	0.730
BYU [4]	0.773	0.820	0.796
DMRZ [4]	0.854	0.863	0.859
dhSegment (retrained CNN)	0.801	**0.945**	0.862
dhSegment [9]	0.826	0.924	0.872
proposed	**0.858**	0.935	**0.895**

In table II, we compare our method with state-of-the-art methods. Even if our method is essentially designed to

[2]https://www-intuidoc.irisa.fr/en/cbad-table-subset/

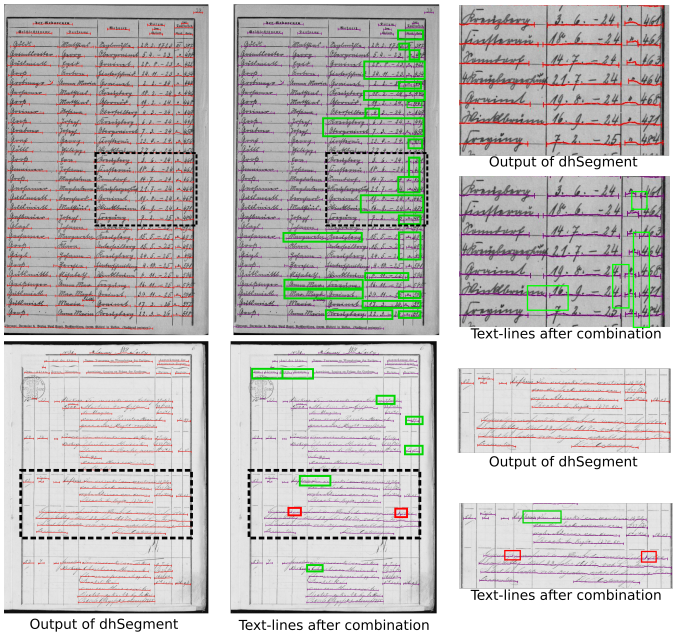

Figure 9. Improvements and errors produced by our combination-(red) output of dhSegment, (purple) lines after our combination, (green) amelioration, (red) error

improve text-line segmentation in tabular structures, it still improves results in a database that does not only contain tabular structures. Compared to state-of-the-art methods, we increase the F-value from 87,2% to **89,5%**. Another method that has not been published yet (but available on arXiv), obtains promising results that outperform ours [5].

V. Conclusion

In this article, we have presented a method that improves baseline detection in handwritten documents containing tables. Our method combines contextual and structural elements in order to interpret the interaction between text and tabular structures. The elements that our method combines are:

- information obtained with different CCN output: hypothetical baselines, begin and end-indicators of baselines, page boundaries,
- structural indicators,
- knowledge about tabular structures to recognize them.

Therefore, our method is able to combine heterogeneous information and this combination depends on the tabular structure to be recognized. Moreover, our method allows to easily integrating other structural knowledge in order to adapt the combination. We showed on a table subset of cBAD competition (track B) that our method is able to improve the F-value from 91.5% to 93.0% compared to a deep learning approach, which obtained the best published score on this competition. On the full test set of the cBAD competition, which contains tabular structures but not only, our method obtains the best results even if it was originally designed for baseline segmentation in the context of tabular structures. Indeed, we increase the F-value from 87.2% to **89.5%**.

References

[1] B. Coüasnon. Dmos, a generic document recognition method: Application to table structure analysis in a general and in a specific way. *International Journal of Document Analysis and Recognition (IJDAR)*, 8(2-3):111–122, Jun 2006.

[2] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, June 2009.

[3] M. Diem, F. Kleber, S. Fiel, T. Grüning, and B. Gatos. Baseline detection in historical documents using convolutional u-nets. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 37–42. IEEE, April 2018.

[4] M. Diem, F. Kleber, B. Gatos S. Fiel, and T. Grüning. cbad: Icdar2017 competition on baseline detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1355–1360. IEEE, 2017.

[5] T. Grüning, G. Leifert, T. Strauß, and R. Labahn. A two-stage method for text line detection in historical documents. In *arXiv preprint arXiv:1802.03345*, 2018.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016.

[7] A. Lemaitre, J. Camillerapp, and B. Coüasnon. Handwritten text segmentation using blurred image. In *Document Recognition and Retrieval XXI*, volume 9021, page 90210D. International Society for Optics and Photonics, March 2014.

[8] I. Leplumey, J. Camillerapp, and C. Queguiner. Kalman filter contributions towards document segmentation. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 765–769. IEEE, August 1995.

[9] S. A Oliveira., B. Seguin, and F. Kaplan. dhsegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 7–12. IEEE, August 2018.

[10] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, page 234–241. Springer, Cham, October 2015.